Advice on total-score reliability issues in psychosomatic measurement

Sijtsma, K.; Emons, W.H.M.

Published in:
Journal of Psychosomatic Research

DOI:
10.1016/j.jpsychores.2010.11.002

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Psychosomatic Research 70 (2011) 565–572

Journal of Psychosomatic Research

Review Article

Advice on total-score reliability issues in psychosomatic measurement

Klaas Sijtsma*, Wilco H.M. Emons

Tilburg University, Tilburg, The Netherlands

Received 24 May 2010; received in revised form 2 November 2010; accepted 4 November 2010

Abstract

Objective: This article addresses three reliability issues that are problematic in the construction of scales intended for use in psychosomatic research, illustrates how these problems may lead to errors, and suggests solutions. **Methods:** We used psychometric results and present five computational studies. The first, third, and fourth studies are based on the generation of artificial data from psychometric models in combination with distributions for scale scores, as is common in psychometric research, whereas the second and fifth studies are analytical. **Results:** The power of Student's *t* test depends more on sample size than on total-score reliability, but reliability must be high when one estimates correlations involving test scores. Short scales often do not allow total scores to be significantly different from a cutoff score. Coefficient alpha is uninformative about the factorial structure of questionnaires and is one of the weakest estimators of total-score reliability. **Conclusions:** The relationship between questionnaire length/reliability and statistical power is complex. Both in research and individual diagnostics, we recommend the use of highly reliable scales so as to reduce the chance of faulty decisions. The conclusion calls for profound statistical research producing hands-on rules for researchers to act upon. Factor analysis should be used to assess the internal consistency of questionnaires. As a reliability estimator, alpha should be replaced by better and readily available methods.

Keywords: Coefficient alpha; Individual decision making; Power analysis; Short scales

Introduction

This article addresses three important but often neglected issues related to total-score reliability of scales intended for use in psychosomatic and other health-related research. These scales measure not only, for example, depression and anxiety [1,2], chronic fatigue [3], and Type D personality [4] but also coping, emotions, and compliance and health-related quality-of-life aspects, such as mobility around the house after surgery [5] and individual feelings and perceptions about people's daily life [6].

The first issue is that researchers too easily accept minimum values for total-score reliability in research settings. We show that such values depend on the research goal and discuss the statistical power of Student's *t* test and estimation of correlations as examples. We draw some tentative conclusions and notice that a finer-grained advice requires a larger-scale statistical investigation.

The second issue is that researchers use short and reliable scales for research and individual decision making more and more. We argue that even for satisfactory total-score reliability, person measurement using short scales may be imprecise, evoking many decision errors [7]. We provide some research results but noticed that this topic also requires large-scale research before definitive advice can be given.

The third issue concerns coefficient alpha [8], which is often used both as an index of internal consistency, meaning the degree to which the items in a questionnaire measure the same attribute, and as an estimator of total-score reliability. We argue that, contrary to common belief, alpha is not an index of internal consistency and further that, as a reliability estimator, it should be replaced by available superior alternatives.

* Corresponding author. Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: +31 13 4663222 or +31 13 4662544; fax: +31 13 4663002.

E-mail address: k.sijtsma@uvt.nl (K. Sijtsma).

0022-3999/10/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jpsychores.2010.11.002

Our purpose is to make the reader aware of these three highly complex issues and the risks involved when they are underestimated in practical scale construction. Underestimation happens when one assumes that (a) a total-score reliability of, say, .6–.8 is sufficient for testing hypotheses about mean scores without ascertaining whether statistical power is high enough; (b) a scale consisting of a small number of high-quality items is reliable enough for individual decision making without checking whether enough decisions are correct; and (*c*) an alpha value of, say, .8 means that the items measure the same attribute without investigating their factorial structure.

This article is technical, but we tried to get our points across using only a few equations. The first two issues are too complex to lead to definitive practical guidelines, and we noticed that more profound statistical research than we can provide here is needed for deriving such guidelines. The third issue concerning coefficient alpha has led to clear-cut conclusions resulting in unambiguous advice. We hope this article stimulates researchers to think hard about their scales, the methodological problems they must resolve, and the psychometric methods they use.

Reliability requirements in research

Measurement instruments may be used in psychosomatic research or for assessing patients' individual behavior. Research applications address, for example, the comparison of means between groups of patients and controls and estimating correlations of total scores with treatment outcomes. Individual assessment refers to the measurement of an individual with the purpose of deciding, for example, whether he or she should receive treatment or whether his or her health condition has improved due to treatment.

Researchers and practitioners have to use their resources efficiently, and a reasonable question is whether short scales may be used instead of longer scales. A longer scale has higher total-score reliability, but perhaps a short scale based on only the highest-quality items leads to little loss of statistical power and estimation precision (research) and reliability (assessing individual performance). In this section, we discuss the consequences of using short scales in research and individual assessment. These constitute the first two issues we cited. First, we discuss the classical test theory (CTT).

CTT, total-score reliability

We assume a questionnaire contains J items and each item has an index number j running from 1 to J and scores running from 0 to m. The total score on the J items is the sum of the item scores, X_j, and defined as $X=\sum_{j=1}^{J} X_j$. If the questionnaire has six rating-scale items with scores running from 0 through 4, then respondents can have total scores, X, running from 0 to 24. The scale length (SL) is the range of possible X scores and equals mJ; here, $SL=24$.

CTT splits total score X in a systematic part or true score T and a random measurement error E, such that $X=T+E$ [9]. Total-score reliability quantifies the degree to which performance in a group is systematic and predictable as opposed to being random and unpredictable. Reliability is defined in two ways, which are mathematically identical.

First, suppose it were possible to readminister the questionnaire a second time under exactly the same circumstances as the first. The reliability of the total score is then technically defined as the correlation between the total scores obtained at the first and second occasions, denoted X and X', and the reliability is the correlation, $\rho_{XX'}$. This definition answers the question that is pervasive in statistics: What would happen if I could do it again?

Second, in a group, the total-score variance can be decomposed into the sum of the true-score variance and the random error variance, $\sigma_X^2=\sigma_T^2+\sigma_E^2$. Reliability is defined as the proportion of total-score variance that is due to true-score variance, $\text{rel}=\sigma_T^2/\sigma_X^2=\sigma_T^2/(\sigma_T^2+\sigma_E^2)$. Higher reliability means a larger share of true-score variance. If it were possible to readminister the questionnaire a second time under exactly the same circumstances as the first, the reliability would remain the same: rel=rel′. The first definition is mathematically identical with the second, so that $\rho_{XX'}=\sigma_T^2/\sigma_X^2=\sigma_{T'}^2/\sigma_{X'}^2$. Thus, the degree to which total scores can be repeated under the same circumstances equals the ratio of true-score variance to total-score variance in both administrations. Because $\sigma_T^2\leq\sigma_X^2$ and because variance is nonnegative, the reliability assumes values between 0 and 1.

An individual's true score T differs for different instruments measuring the same attribute. Assume we have a four-item questionnaire in which each item has a rating scale with scores 0,…,4, so that total score X has values 0,…,16. Let John have a true score $T_{\text{John}}=6$ and Mary a true score $T_{\text{Mary}}=10$. We double the length of the questionnaire by another four-item set, which is parallel with the first four-item set (parallelism is: the two item sets are psychometrically identical but have different contents [9]); then, the true scores also double: $T_{\text{John}}=12$ and $T_{\text{Mary}}=20$, and so does their difference. In real life, subsets of items are never parallel and true scores change without neatly doubling.

Research instrument, reliability, and sample size

Testing group-means difference

We considered the following research question: Suppose a group of patients and a group of controls have different mean depression levels. A researcher has constructed a research instrument meant for measuring depression. He or she intends to compare the two groups using the instrument. For this purpose, the researcher has collected data from both groups and used Student's t test to test whether the mean instrument scores in the two groups are equal (null hypothesis) or not. How does the total-score reliability affect the statistical test results? This is a complex problem (e.g.,

Refs. [10] and [11]) to which no easy answers are available, but here we provide tentative suggestions.

In general, as reliability increases, the share of the true-score variance in the total-score variance is larger and one expects group differences in mean true scores to become larger. Hence, one also expects differences between mean total scores to become larger. Thus, a more reliable questionnaire is expected to better pick up existing differences in depression levels between groups, and Student's t test may signal a smaller difference already as significant. What makes this problem complex is that reliability may increase in different ways: by keeping total-score variance constant while letting true-score variance increase, by keeping true-score variance constant while letting error variance decrease, by letting the number of items grow by adding items of the same psychometric quality, by starting with a small set of the best items and then adding items of lower quality, and so on. The point is that the manipulation of many questionnaire and item properties affects reliability and it is rather unpredictable how Student's t test reacts to this manipulation. We briefly consider the power of Student's t test for increasing number of items (J), also called questionnaire length.

Student's t statistic for independent samples with unequal sizes and variances equals (P=patients, C=controls, n=subgroup size),

$$t = \frac{\overline{X}_{\mathrm{P}} - \overline{X}_{\mathrm{C}}}{S_{\overline{X}_{\mathrm{P}} - \overline{X}_{\mathrm{C}}}},\ S_{\overline{X}_{\mathrm{P}} - \overline{X}_{\mathrm{C}}} = \sqrt{\frac{S_{\mathrm{P}}^2}{n_{\mathrm{P}}} + \frac{S_{\mathrm{C}}^2}{n_{\mathrm{C}}}}.$$

Increasing the questionnaire length, and hence the reliability, affects the group means and the S.E. values in the t statistic, but it is difficult to predict what happens to the t test's power.

We conducted a computational study using simulated data for five-point rating-scale items to investigate the power of the t test as questionnaire length increases and hence reliability increases. Because data are simulated, the alternative hypothesis is known (details about the study can be obtained from the second author). Table 1 shows that as questionnaire length increased from 5 to 20 items, Cronbach's alpha increased from .70 to .90. The difference between the group means (numerator of t statistic) and the corresponding S.E. (denominator) both increased, and statistic t and standardized effect size d increased but only little. P values decreased, and at the 5% level, we found a significant result for J=15 and J=20 but not for J=5 and J=10. The power of the t test increased little as questionnaire length increased but greatly as sample size increased for fixed reliability (Table 1, three rightmost columns). We conclude that for testing group-means difference, more power is gained by increasing sample size than by increasing reliability, say, from .80 to .90. Does this result hold for any statistical analysis? The answer is *no*, as the next example shows.

Table 1
Questionnaire length and Student's t-test results

							Power		
J	alpha	$\overline{X}_P - \overline{X}_C$	S.E.	t	P	d	N=50	N=100	N=300
5	.70	−0.99	0.55	−1.80	.075	−0.26	.26	.43	.87
10	.81	−1.96	1.01	−1.94	.055	−0.28	.29	.50	.94
15	.87	−2.96	1.47	−2.01	.047	−0.29	.28	.52	.94
20	.90	−3.94	1.93	−2.04	.044	−0.29	.31	.53	.94

Shown for four levels of questionnaire length (J) are the coefficient alpha, difference in means ($\overline{X}_P - \overline{X}_C$), S.E. (denominator t statistic), t statistic, P value, and standardized effect size (d) for N=100 and small effect size, and the power of independent samples t test for varying sample sizes (N) for small effect size and a nominal significance level of 5%. For each value of J, 1000 data sets were simulated. Table entries are means across 1000 data sets.

Estimating correlations with other variables

Suppose one wishes to know the correlation between the depression total score X and a criterion score Y indicating suitability for a particular treatment, which is also represented by a fallible score. Then, to learn how the reliabilities of both measurements affect their correlation, ρ_{XY}, the attenuation correction formula [9] from CTT can be written as

$$\rho_{XY} = \sqrt{\rho_{XX'}\rho_{YY'}} \times \rho_{T_X T_Y},$$

with T_X and T_Y denoting the true-score components of total scores X and Y, respectively.

Clearly, the correlation between the true scores is of interest, but in practice, one only has access to the fallible total scores. The equation shows that ρ_{XY} is attenuated by measurement error: the lower the reliabilities, the lower the correlation between the observable scores and the more ρ_{XY} is a distorted estimate of $\rho_{T_X T_Y}$. Arbitrarily, assume that $\rho_{T_X T_Y}$=.6, which is the correlation of interest; then, only if $\rho_{XX'}$=$\rho_{YY'}$=1 does one find that ρ_{XY}=$\rho_{T_X T_Y}$. In practice, perfect reliability is too good to be true. Lowering one or both reliabilities yields a number $\sqrt{\rho_{XX'}\rho_{YY'}}$ between 0 and 1; hence, ρ_{XY} is always lower than $\rho_{T_X T_Y}$. For example, take $\rho_{XX'}$=$\rho_{YY'}$=.8, then ρ_{XY}=.8×.6=.48. For $\rho_{XX'}$=.6 and $\rho_{YY'}$=.8, one finds $\rho_{XY} = \sqrt{.48} \times .6 \approx .42$.

We conclude that the lower the reliabilities, the lower the correlation between the two total scores and hence the more distorted the correlation of interest. A larger sample does not help; the same distortion is estimated but with more precision. The effect of larger sample size on the precision is illustrated using another computational study.

Table 2 shows for $\rho_{T_X T_Y}$=.6 and varying combinations of reliabilities of X and Y the resulting correlation ρ_{XY}. We took each of the 10 population values of ρ_{XY} as correlations found in real research (i.e., as sample values r_{XY}), and then for increasing sample size N, we computed confidence intervals (CIs) for the population correlation ρ_{XY}. For this purpose, we used the Fisher

Table 2
Correlation between scale scores X and Y (ρ_{XY}) and corresponding 95% CIs for 10 combinations of reliability of X ($\rho_{XX'}$) and Y ($\rho_{YY'}$) and five levels of N, given a true-score correlation ($\rho_{T_XT_Y}$) of .6

			95% CI				
$\rho_{XX'}$	$\rho_{YY'}$	ρ_{XY}	N=50	N=100	N=200	N=500	N=1000
.4	.4	.24	−.04 to .49	.05–.42	.10–.37	.16–.32	.18–.30
.4	.6	.29	.01–.53	.10–.46	.16–.41	.21–.37	.23–.35
.4	.8	.34	.07–.56	.15–.50	.21–.46	.26–.42	.28–.39
.4	.9	.36	.09–.58	.18–.52	.23–.48	.28–.43	.30–.41
.6	.6	.36	.09–.58	.18–.52	.23–.48	.28–.43	.30–.41
.6	.8	.42	.16–.63	.24–.57	.30–.53	.35–.49	.37–.47
.6	.9	.44	.18–.64	.27–.59	.32–.55	.37–.51	.39–.49
.8	.8	.48	.23–.67	.31–.62	.37–.58	.41–.54	.43–.53
.8	.9	.51	.27–.69	.35–.64	.40–.61	.44–.57	.46–.55
.9	.9	.54	.31–.71	.38–.67	.43–.63	.47–.60	.49–.58

z-transformation (writing r instead of r_{XY} and ln for the natural logarithm),

$$z = \frac{1}{2} ln \frac{1+r}{1-r},$$

to normalize the sampling distribution of r. The distribution of z has S.E.,

$$S.E.(z) = \frac{1}{\sqrt{N-3}}.$$

Using S.E.(z), we computed 95% CIs, z±1.96S.E.(z), and then used the inverse transformation (writing exp for the exponential function),

$$\frac{exp(2z)-1}{exp(2z)+1},$$

to transform the lower and upper bounds of the interval back to the scale of r. This produced the well-known asymmetrical CIs for the correlation ρ_{XY} shown in Table 2.

Table 2 shows that for each correlation (third column), the way the CIs are estimated maintains the same midpoint based on the correlation (but not identical with it, due to asymmetry) but produces a narrower interval as sample size N grows. Hence, by using the sample correlation and the sample size, one estimates the same distorted parameter ρ_{XY} but with more precision.

We do not recommend using the formula for attenuation correction the other way around by inserting the reliabilities and the correlation between total scores X and Y and then estimating the correlation between the true scores. This estimate is based on fallible total scores, and the question on what would really happen with the correlation between X and Y when both measures were improved by adding more items or replacing malfunctioning items by better items remains unanswered. The attenuation correction gives a prediction, not a value based on real instrument construction and real data collected by means of these improved instruments.

To summarize, when group means are compared, a higher reliability increases statistical power only little and a larger sample size is more effective. When the correlation of a total score with another total score is estimated, imperfect reliability distorts the desired outcome. Fleshing out the minimum-reliability problem for all kinds of hypotheses and estimation problems requires profound statistical research, which is beyond the scope of this study. Based on the present state of knowledge, rules of thumb for total-score reliability are difficult to set up: Sometimes reliability seems to be subordinate to sample size, and sometimes it is dominant, nearly independent of sample size; hence, the general problem is complex and in need of further research.

Reliability requirements in individual assessment

For determining the usefulness of total scores in making decisions about individuals, size N of the sample in which the psychometric qualities of the scale were ascertained does not play a role. What does play a role are the individual's J item scores. The greater J, the more information is available about the individual's true scores and the more precise is the individual's true-score estimate. In this section, we discuss first the precision of total score X, then the use of total score X for classifying patients into one of two clinical groups, and finally the use of total-score change for assessing individual change due to therapy.

Total-score precision

We use the total score X to estimate T. Let the estimated value be denoted by $\hat{T}$, then $\hat{T}=X$. If John has a score of 19 points, we take this to be his true-score estimate. This is what almost all applied researchers do—that is, they use X, but more sophisticated methods would lead to the same conclusion. It is of great interest to know how precise the estimate $\hat{T}=X$ is. For this purpose, psychometricians use the CI for T. Very few researchers use CIs and instead take X as if it were the true score, tacitly assuming $\rho_{XX'}=1$; but then why bother about total-score reliability at all?

To determine a CI, we need a statistical model for drawing samples—in this case, a total score from a distribution of total scores obtained with the same instrument. Ideally, for each individual, such a distribution would be available, but in real life, each individual produces only one total score. In the absence of such individual distributions, the common practice is to use the S.D. of the measurement error for the whole group, σ_E, and assume it also is the S.D. of an individual's hypothetical distribution of total scores. S.D. σ_E is called the standard measurement error (SME), and equals,

$$SME = \sigma_E = \sigma_X \sqrt{1-\rho_{XX'}}.$$

The mean of an individual's hypothetical distribution of total scores is his or her true score; for John, that is $\mu_{John}=T_{John}$.

Furthermore, it is assumed that John's distribution is normal; in statistical notation, $X_{\text{John}} \sim N(T_{\text{John}}, \sigma_E^2)$. The SME is estimated by inserting the S.D. of the total score in the group, S_X, which estimates σ_X, and an estimate $r_{XX'}$ for the total-score reliability $\rho_{XX'}$, for example, Cronbach's alpha, so that,

$$S_E = S_X \sqrt{1 - r_{XX'}} .$$

A $(1-\alpha)\times 100\%$ CI is computed as $\hat{T}_{\text{John}} \pm z_{1/2\alpha} S_E$, where $z_{1/2\alpha}$ is the normal deviate that corresponds to the area under the normal curve for the CI. For a 95% CI, in a table for the standard normal distribution, one finds $z_{0.025}=1.96$; for a 90% CI, one finds $z_{0.050}=1.645$. With the use of S_E and a choice for $z_{1/2\alpha}$, a CI with the same width can be computed for each true-score estimate.

Individual assessment, reliability, and number of items

Classification using the total score

Suppose one uses total score X for assigning people to one of two groups. The decision rule is: if $X<X_c$ (where c is cutoff), assign to the no treatment condition, and if $X \geq X_c$, assign to the treatment condition. Then, one has to check whether the cutoff falls in the CI for a particular individual. If it does, then total score X is not significantly different from the cutoff and a reliable decision cannot be made; if it does not, then lower total scores support the nontreatment decision and higher total scores support the treatment decision.

In practice, reliability is often used in a somewhat ritualistic way. For example, if reliability is, say, at least .8, it is assumed that the scale is precise enough for making individual decisions, and total scores are used as if they were true scores. However, .8 is smaller than 1, and as long as a scale has imperfect reliability, one should test whether the true score equals the cutoff. Next, we demonstrate how reliability is related to SMEs, CIs, and SL. This yields surprising results, urging researchers and practitioners to be cautious, especially when they use short scales.

We did an experiment using the item response theory model known as the graded response model [12] to define the psychometric properties of a set of J items comprising a questionnaire and combined the artificial item set with a distribution of the attribute scores in a group of interest. Next, we generated data for N individuals who responded to J items with five ordered answer categories yielding scores 0,…,4, analyzed the data set, and repeated the experiment for different numbers of items and varying item and questionnaire properties. Items had their highest discrimination power around cutoff X_c, and total-score reliability for the longest questionnaire equaled .9. Shorter scales consisted of item subsets from the longer questionnaire. (Please consult the first author for more information.)

Table 3 shows total-score reliability, SME, 90% and 95% CIs for true score T, which is located exactly at half the scale, and the ratio of CI and SL (CI/SL). The question is whether true-score estimate $\hat{T}$ differs significantly from cutoff X_c, which lies to the right of T at 60% of the scale. Only if $\hat{T}$ is significantly smaller than X_c is the individual admitted to nontreatment. Table 3 shows (from bottom to top) that reliability and SME decreased as questionnaire length J decreased. Nevertheless, many researchers would consider the reliability sufficient for short scales, and their view seems to be supported by smaller SMEs resulting in narrower CIs; hence, it seems that the true score can be estimated with greater precision. However, given the smaller number of items and the corresponding loss of information, this cannot be true. The catch is that SL has decreased faster than the CIs, so that CI/SL at least has doubled. Fig. 1 shows that the cutoff has moved into the CI for small J and that it can no longer be concluded that the individual should be assigned to nontreatment.

Table 3
SL, coefficient alpha, SME, 90% and 95% CIs for the true score (T) half way the scale, and CI/SL for each CI for four levels of questionnaire length (J)

J	SL	alpha	SME	CI_{90}	CI/SL	CI_{95}	CI/SL
5	20	.70	1.83	6.98–13.02	.30	6.41–13.59	.36
10	40	.80	2.24	16.32–23.68	.18	15.62–24.38	.22
15	60	.87	2.76	25.47–34.51	.15	24.60–35.40	.18
20	80	.90	3.18	34.77–45.23	.13	33.77–46.23	.16

Results for the whole table are based on a data set of 1000 simulated item-score vectors using five-point rating-scale items (i.e., m=4).

The example shows that, in spite of good reliability based on good-quality items, short tests can deceive in a terrible way. Emons et al. [7] have taken this topic further and demonstrated that small numbers of good-quality items easily classify respondents in the wrong condition. The problem is small SL, which leaves total scores less room to be different from cutoffs even when CIs are narrow. Long(er) scales reduce percentages of classification errors considerably.

Assessing individual change

We considered effect of therapy as reflected in total-score change. Individual change is the difference between an individual's total score after treatment, X_2, and his or her total score before treatment, X_1. The reliable change (RC) index [13] is defined as follows:

$$RC = \frac{X_2 - X_1}{S_{diff}},$$

where $S_{diff} = \sqrt{2S_E^2}$ (S_E is the sample SME) under the hull hypothesis of no change. The use of the RC statistic assumes normally distributed measurement errors. One may test (e.g., at the 5% level) whether observed change in the expected direction of recovery ($X_2-X_1>0$) is significant ($RC>1.645$) or whether observed change in any direction, recovery or deterioration ($X_2-X_1\neq 0$), is significant ($|RC|>1.96$).

For the same computational setup that underlay the results in Table 3, Table 4 shows power results for improvement

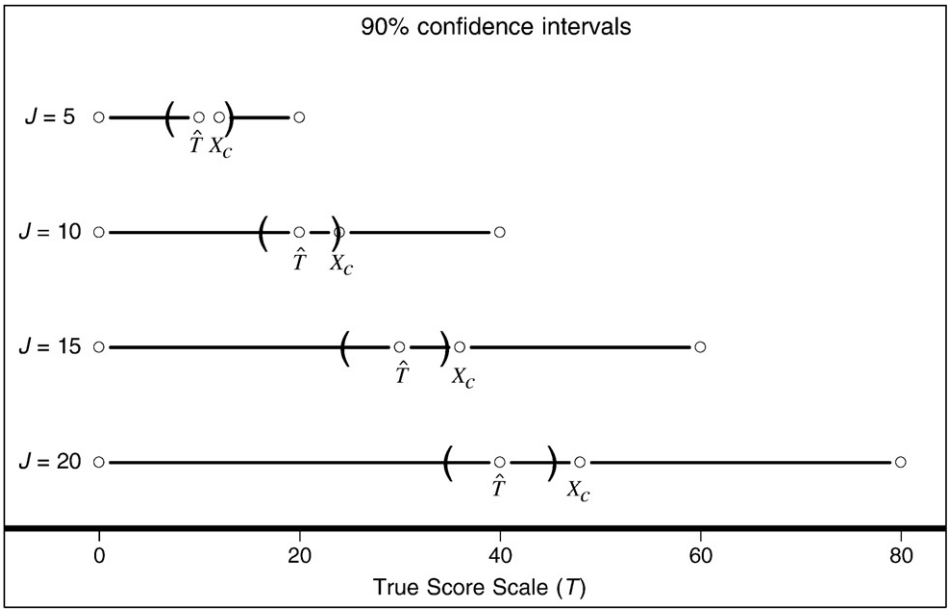

Fig. 1. Ninety percent CIs for a true score located at 50% of the scale, for different questionnaire lengths, and a cutoff score that is at 60% of the scale.

(one-tailed test) and change (two-tailed test). First, in identical circumstances, the one-tailed test always has more power than the two-tailed test. Second, power is only substantial (say, >.60) if the effect sizes are large (1 S.D. of the true person-distribution) or very large (2 S.D.s). For small effect size (0.25 S.D.), even 20-item tests fail to produce a powerful one-sided test (power=.37); for medium effect size (0.5 S.D.), only tests in excess of 10 items have sufficient power.

Again, questionnaire length proves to be important in individual decision making. This should come as no surprise as change assessment also involves classification: Did the individual improve or change significantly or not? Jacobson and Truax [13] rightly argued that in addition to statistically significant change, an individual must show clinically significant change in order to be recovered. This requires the individual to cross a cutoff score X_c that can be determined in different ways depending on the dysfunctional and functional total-score distributions (Ref. [13], p. 13). Jacobson and Truax ([13], p. 16) also suggested statistically testing whether a posttreatment score X_2 is significantly greater than cutoff X_c. In our computational study, we only considered the power of the RC statistic, but if we would also require individuals having significant RC values to show clinical significance and to have posttreatment scores significantly larger than X_c, the power of the whole procedure would be smaller compared with the results in Table 4; also see Jacobson and Truax [13] for a discussion of this issue.

Coefficient alpha: internal consistency and reliability estimation

Coefficient alpha is often used as an index for internal consistency. The literature is not explicit about the meaning of the term, but *internal consistency* is usually taken to mean that the items constituting the questionnaire measure the same attribute [14,15]. A high value of alpha is then interpreted to mean that the items measure this attribute. However, several authors [15–17] have argued convincingly that there is no relationship between values of alpha and the number of factors underlying the item scores. Thus, they argued that both a one-factorial questionnaire and a multifactorial questionnaire could

Table 4
Power to detect treatment effects using Jacobson and Truax' RC index and significance level of 5% for four levels of individual clinical change

			Power of one-tailed test				Power of two-tailed test			
			Individual change				Individual change			
J	alpha	SME of difference scores	Small	Medium	Large	Very large	Small	Medium	Large	Very large
5	.70	2.96	.19	.39	.76	.95	.12	.27	.65	.91
10	.82	4.12	.26	.58	.96	.99	.14	.49	.90	.98
15	.87	5.06	.29	.68	.98	.99	.18	.61	.97	.99
20	.90	5.84	.37	.80	.99	1.00	.26	.70	.98	.99

Results in each cell are based on 50,000 newly simulated difference scores.

Table 5
Hypothetical covariance matrices for one-factorial questionnaire data and two-factorial questionnaire data, both with alpha=.56

	One-factorial questionnaire				Two-factorial questionnaire			
	1	2	3	4	1	2	3	4
1	.25	.06	.06	.06	.25	.16	.01	.01
2	.06	.25	.06	.06	.16	.25	.01	.01
3	.06	.06	.25	.06	.01	.01	.25	.16
4	.06	.06	.06	.25	.01	.01	.16	.25

have either low or high alpha values. Rather than repeating their results, we call attention to the following line of reasoning, which makes the same point.

The information about the factorial composition of a set of items comprising a questionnaire is in the structure of the correlations between the items. For convenience, we use the inter-item covariances instead of the inter-item correlations. Roughly speaking, if all inter-item covariances are the same (Table 5, left panel), one factor can describe this inter-item covariance structure; if there are two clusters of items such that items within the same cluster covary highly and items from different clusters covary lowly (Table 5, right panel), then two factors can describe this two-cluster inter-item covariance structure, and so on. Now, if alpha were an index for internal consistency, meaning that one factor explains the inter-item covariance structure (i.e., the items measure the same attribute), only one-factorial inter-item covariance structures must produce high alpha values. However, this is not true [15], and a look at the equation for alpha explains why.

Let the inter-item covariance between items j and k be denoted σ_{jk}, and let the mean value of the inter-item covariances between all $J(J-1)$ item pairs (in Table 5, 12 pairs) in the questionnaire be denoted $\bar{\sigma}$, then we can write coefficient alpha as follows:

$$\text{alpha} = \frac{J^2\bar{\sigma}}{\sigma_X^2}.$$

Thus, alpha depends on the mean of the inter-item covariances but not on the individual inter-item covariances. Since the information about the factorial structure of the J items is in the structure of the $J(J-1)$ inter-item covariances and since this information is lost when only the mean of these covariances is available, alpha cannot provide information about the factorial item structure. For example, for both examples in Table 4, one finds that $\bar{\sigma}=.06$ (and that $J=4$ and $\sigma_X^2=1.72$), so that alpha=.56. Hence, alpha is uninformative about the factorial structure and is misleading as an index of internal consistency.

The second use of alpha, as an estimator of total-score reliability, is correct but can be somewhat unfortunate, as we explain next. Researchers often do not seem to realize that alpha is only a lower bound to the reliability—that is, alpha$\leq\rho_{XX'}$ (e.g., Ref. [9]), and that much better estimators are available. Sijtsma [15] recommends using Guttman's λ_2 [18], which, similar to alpha, is available in IBM SPSS Statistics 18, or the greatest lower bound (GLB; [19]), which is the lower bound closest to reliability $\rho_{XX'}$. It is worth noticing that alpha$\leq\lambda_2\leq$GLB$\leq\rho_{XX'}$. For example, for an eight-item questionnaire on coping strategies, Sijtsma [15] reported alpha=.778, λ_2=.785, and GLB=.852. GLB is available in EQS (http://www.mvsoft.com/eqs60.htm; e.g., Ref. [20]).

One could argue that all a lower bound, such as alpha, does wrong is underestimate total-score reliability and that a conservative reliability estimate may even spur researchers to construct a questionnaire that is more reliable than they think. On the other hand, using a reliability estimate that improves upon alpha provides a more realistic reliability estimate and produces narrower CIs for individual decision making. Using better reliability estimates alleviates the need for longer scales at least a little.

Alpha, λ_2, and GLB can be estimated irrespective of the factorial structure of the items; they simply are not informative about it. However, researchers usually want their questionnaires to measure one attribute. Thus, it is advisable to investigate the factorial structure of the items and perhaps remove one or two items that are poor indicators of the dominant factor prior to estimating total-score reliability for the remaining items. Structural equation modeling also provides methods for reliability estimation that may yield better results than alpha, provided the underlying factor model describes the data well [21]. This factor model may contain, for example, second-order factors in addition to a dominant factor.

Conclusions and recommendations

We conclude by providing a couple of take-home messages for instrument constructors, researchers using instruments, and practitioners:

- Higher reliability produces more power for testing differences between means, but a larger sample is more effective. Also, higher reliability reduces the underestimation of the strength of relationships, but sample size has no effect. More research into similar power and estimation problems is badly needed. In general, we recommend using longer scales having a highly reliable total score.
- Statistically, the number of items in a scale is the sample of information available for estimating individuals' attribute levels. Short scales tend to result in faulty decision making. Our advice is to use many items if the decision is important.
- Longer scales that include the items from the short-scale version do everything technically better than the short scale, but they increase the burden on the patient and require more administration time. Alternatively, the researcher may consider using many sources of

information, among them several short scales for different attributes (e.g., Ref. [22]).

- Coefficient alpha is not an index for internal consistency of the questionnaire. Internal consistency must be investigated using methods for dimensionality assessment, such as factor analysis.
- Coefficient alpha should be replaced by better reliability estimation methods, such as λ_2 and the GLB.

Acknowledgments

We thank Marrie H. J. Bekker and Susanne S. Pedersen for providing critical comments on a previous draft of this article.

References

[1] Bjelland I, Dahl AA, Haugh TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res 2002;55:69–77.
[2] Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
[3] Michielsen HJ, De Vries J, Van Heck GL, Van de Vijver FJR, Sijtsma K. Examination of the dimensionality of fatigue: the construction of the Fatigue Assessment Scale (FAS). Eur J for Psychol Assess 2004;20: 39–48.
[4] Denollet J. DS14: standard assessment of negative affectivity, social inhibition, and type D personality. Psychosom Med 2005; 67:89–97.
[5] Roorda LD, Roebroeck ME, Van Tilburg T, Molenaar IW, Lankhorst GJ, Bouter LM, and the Measuring Mobility Studying Group. Measuring activity limitations in walking: development of a hierarchical scale for patients with lower-extremity disorders who live at home. Arch Phys Med Rehabil 2005;86:2277–83.
[6] The WHOQoL Group. Development of the World Health Organization WHOQOL-Bref QoL assessment. Psychol Med 1998;28:551–9.
[7] Emons WHM, Sijtsma K, Meijer RR. On the consistency of individual classification using short scales. Psychol Methods 2007;12:105–20.
[8] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
[9] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.
[10] Levin JR. Note on the relation between the power of a significance test and the reliability of the measuring instrument. Multivariate Behav Res 1986;21:255–61.
[11] Nicewander WA, Price JM. Reliability of measurement and the power of significance tests. Psychol Bull 1983;94:524–33.
[12] Samejima F. Graded response model. In: Van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer-Verlag, 1997. p. 85–100.
[13] Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–9.
[14] Sijtsma K. Correcting fallacies in validity, reliability, and classification. Int J Test 2009;9:167–94.
[15] Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 2009;74:107–20.
[16] Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78:98–104.
[17] Schmitt N. Uses and abuses of coefficient alpha. Psychol Assess 1996; 8:350–3.
[18] Guttman L. A basis for analyzing test–retest reliability. Psychometrika 1945;10:255–82.
[19] Bentler PA, Woodward JA. Inequalities among lower bounds to reliability: with applications to test construction and factor analysis. Psychometrika 1980;45:249–67.
[20] Bentler PA. EQS structural equations program manual. MultivariateEncino, CA: Software, Inc, 1995.
[21] Yang Y, Green SB. A note on structural equation modeling estimates of reliability. Struct Equation Model 2010;17:66–81.
[22] Egberink IJL, Meijer RR. An IRT analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. Assessment [in press].